DOCUMENT RESUME

ED 191 911 TM 800 535

AUTHOR Schwarzmueller, E. Beth: And Others
TITLE Statistical Procedures in Evaluation Models for Selection-Biased Populations.
PUB DATE Apr 80
NOTE 38p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cohort Analysis: *Mathematical Models: Program Evaluation: *Quasiexperimental Design: *Regression (Statistics): Sampling: *Statistical Bias: *Statistical Significance
IDENTIFIERS *Regression Discontinuity Model: *Regression Projection Model: Simultaneous Region Procedure

ABSTRACT

Statistical procedures of the regression discontinuity and regression projection designs were compared to determine which design was more able to detect treatment effect in remedial or gifted programs. Neither test was more precise in all situations. Tables list the more precise test under specific conditions. The regression projection model produced a more precise test when the sample sizes of the two groups were equal or nearly equal. The regression discontinuity model was more precise when the sample size of the treatment group was much smaller than that of the comparison group. Both models led to erroneous conclusions about program impact when a treatment-selection variable interaction existed. For this situation, a simultaneous region procedure based on Scheffe intervals is suggested. This provides a test of the treatment effect at all levels of the selection variable, allowing the user to infer for which individuals--none, some, or all--an effect exists. A derivation of the simultaneous region procedure and a practical example demonstrating its use are presented. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED191911

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

STATISTICAL PROCEDURES IN EVALUATION MODELS FOR SELECTION-BIASED POPULATIONS

by

E. Beth Schwarzmueller
C & S National Bank
Atlanta, Georgia

and

John H. Neel, Martin D. Fraser, William L. Curlette,
and William F. Stallings
Georgia State University
Atlanta, Georgia

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
E. B. Schwarzmueller
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper Presented at the
American Educational Research Association
Annual Meeting
Boston, Massachusetts
April, 1980

*This paper is based on the first author's doctoral dissertation, which was completed in partial fulfillment of the requirements for the Ph.D. degree in Educational Research and Measurement at Georgia State University, June, 1979.

TM 800 535

Introduction

In many educational situations, special treatments or programs are offered to a select group of students, those who are more deserving or most in need. For example, programs for gifted children are usually offered to all children who score above a certain cutoff score on some measure of intelligence. One educational problem is the evaluation of the impact or effects of these special programs. All children eligible for the program are allowed to participate, and thus there is no population to serve as a comparison group. That is, there is no identifiable control group, a requirement of evaluations using true experimental designs as described by Campbell and Stanley (1963). In the absence of true experimental design, evaluations of the worth of the programs are often based on descriptive statistics and/or personal opinion. Evaluations directed toward determining program impact should be judged, at least in part, on objective evidence. For this reason, evaluations of educational programs require experimental design that will provide this objective evidence (Anderson & Ball, 1978).

Two evaluation models using regression techniques have been proposed for making inferences about program impact in situations where there is a selection bias. The first of these is the regression discontinuity design proposed by Thistlethwaite and Campbell (1960) as a quasi-experimental design for use when selection is confounded with treatment. This technique involves regressing

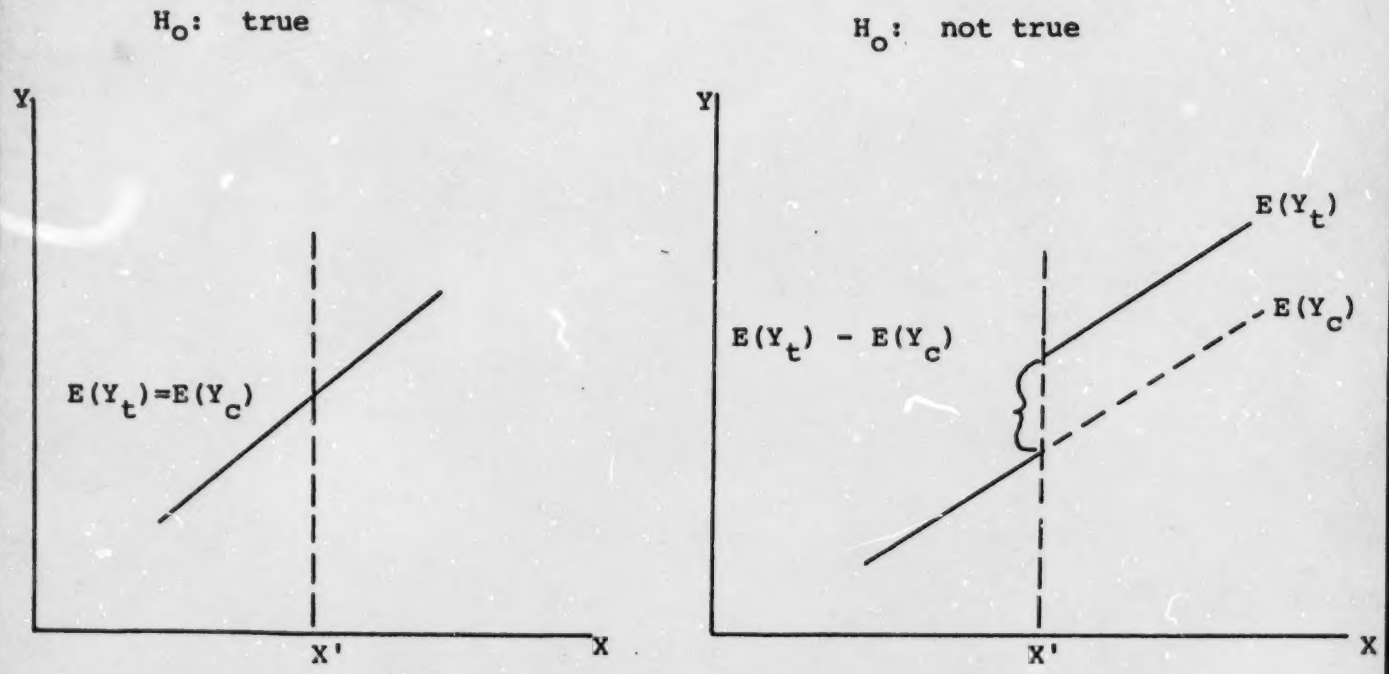
the dependent variable on the variable used for determining placement separately for each group, those selected and those not selected. If the null hypothesis of no program impact is true, the two regression models predict the same score on the dependent variable for each value of the selection variable. If the program has had an effect, there is discontinuity in the regression models at the cutoff point. Thistlethwaite and Campbell suggested that a sharp discontinuity at the cutoff point was evidence of a treatment effect (see Figure 1). The appropriate statistical test for evaluating the statistical significance of $\hat{Y}_t - \hat{Y}_c$ for a given X was given by Walker and Lev (1953, p. 400).

$$t_{N-4} = \frac{\hat{Y}_t - \hat{Y}_c}{s_{\hat{Y}_t - \hat{Y}_c}} \tag{1}$$

where $\hat{Y}_t = a_t + b_t X$, $\hat{Y}_c = a_c + b_c X$, a and b are the usual least squares estimates of intercept and slope, and

$$s_{\hat{Y}_t - \hat{Y}_c} = \sqrt{\frac{\left(\sum_{j=1}^{2}\left(\sum_{i=1}^{N_j} y_i - \frac{\left(\sum_{i=1}^{N_j} x_i y_i\right)^2}{\sum_{i=1}^{N_j} x_i^2}\right)\right)\left(\frac{N}{N_t \cdot N_c} + \sum_{j=1}^{2} \frac{(X' - \bar{X}_j)^2}{\sum_{i=1}^{N_j} x_i^2}\right)}{(N-4)}} \tag{2}$$

j = group, N_j = number in each group, X' = cutoff score, $N = N_t + N_c$. Campbell included this design in several works on experimental design (Campbell, 1969; Campbell & Stanley, 1963; Rieken & Boruch, 1974). Sween (1972) expanded on the regression discontinuity



X = selection variable; Y = dependent variable; X' = cutoff score

$E(Y_i) = \alpha_i + \beta_i X$ $\quad$ $E(Y_t) - E(Y_c)$ = treatment effect

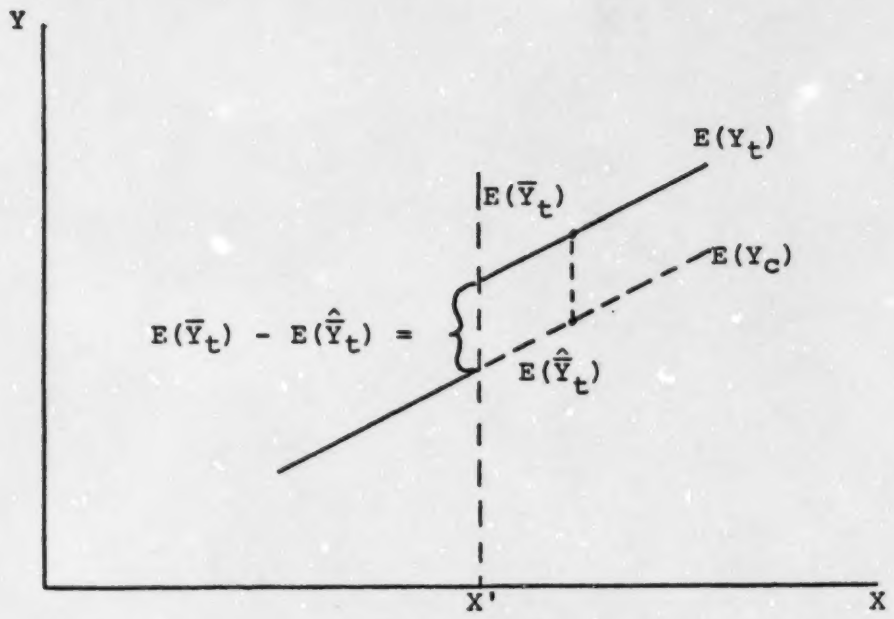
Figure 1. Regression discontinuity design for null and alternative conditions.

design, stressing the importance of identifying the underlying regression. Using a linear model when a quadratic or cubic model better fit the data produces pseudo effects. She, therefore, recommended fitting at least first, second, and third order regression models to the data to determine the best fit. The test of the statistical significance suggested was still a test of $\hat{Y}_t - \hat{Y}_c$ at the cutoff point with a slight modification of the standard error of the difference for the higher order regression models.

The second evaluation model using regression techniques is the regression projection model. This model has been proposed for evaluating the impact of supplementary instructional services, specifically Title I projects, in which selection to treatment is determined by need (Tallmadge & Wood, 1978).

Again, the dependent variable scores are regressed on the selection variable scores for each group separately. In the regression projection model, effects of the program are measured by comparing the selected group's mean performance on the posttest with an estimate of what it would be without treatment, the predicted score corresponding to the selected group's mean pretest score (see Figure 2). A significant difference between the observed mean of the treatment group ($\bar{Y}_t$) and the predicted mean ($\hat{\bar{Y}}_t$) based on the regression line of the comparison group is evidence of a treatment or program impact. Tallmadge and Horst (1976, p. 64) suggested this statistical test for determining the significance of the difference:

H_o: not true



X = selection variable; Y = dependent variable

X' = cutoff score

$E(\bar{Y}_t) - E(\hat{\bar{Y}}_t)$ = treatment effect

$\hat{\bar{Y}}_t = \bar{Y}_c + b_c(\bar{X}_t - \bar{X}_c)$

Figure 2. Regression projection model for alternative condition.

$$t_{N-3} = \sqrt{\frac{(N_t/N)^2 \ (\bar{Y}_t - \hat{\bar{Y}}_t)^2 (N - 3)}{b_c^2 \bar{s}_X^2 + \bar{s}_Y^2 - 2b_c \bar{b} \bar{s}_X^2 + (N_t N_c / N^2)(\bar{Y}_t - \hat{\bar{Y}}_t)^2}} \qquad (3)$$

where

$$N = N_t + N_c$$

$$\bar{s}_X^2 = \frac{N_t}{N} s_{X_t}^2 + \frac{N_c}{N} s_{X_c}^2$$

$$\bar{s}_Y = \frac{N_t}{N} s_{Y_t}^2 + \frac{N_c}{N} s_{Y_c}^2$$

and

$$\bar{b} = \frac{N_t}{N} b_t + \frac{N_c}{N} b_c \ .$$

Both of these evaluation models assume the effects of the treatment or program are constant over all values of the selection variable. If not, there would not only be differences in the intercepts of the regression lines, but also in the slopes. A difference in slopes implies an interaction between the treatment and selection variable, or a nonlinear relationship between the dependent and selection variables, or both. Researchers may have assumed constant treatment effects because it is difficult to distinguish between these interactions and nonlinear relationships. However, there may be situations in which information about the underlying relationship is available. In that case, there may be other techniques for determining regions of significance which would be more useful in evaluating the effects of the program.

In this study, the regression discontinuity and regression projection models were compared to determine each model's ability

to provide objective evidence for making decisions about the worth of educational programs. In particular, the following issues were addressed. Are the given statistical tests appropriate? If so, which model provides the more powerful or precise test for detecting treatment or program effect? If they are not appropriate, is there a statistical test which can be used to determine a region of significance--that is, a region such that, for all participants whose scores on the selection variable fall within that region, there is a significant difference between the two regression lines? It could then be concluded that there is a program impact for the individuals whose scores fall in this region.

Comparison

Both the regression discontinuity and regression projection evaluation models assume that all subjects in a sample are measured on a selection variable. All subjects who score beyond the cutoff (above or below, depending upon the situation) become participants in the supplementary program or treatment group. The remaining subjects serve as the comparison group. Any exception to this assignment rule invalidates the results and should be excluded from the data analysis.

In both evaluation models, program effectiveness is determined by testing all subjects in the sample on some measure of the criterion or dependent variable. The use of these models assumes a linear relationship between the dependent and selection variables and homoscedasticity of the dependent variable at each value of the selection variable. The dependent variable scores are

regressed on the selection variable scores for each group separately. The statistical test for each model provides evidence about the presence of treatment effect.

As given in Equations 1 and 2, the appropriate test for the regression discontinuity model is

$$t_{N-4} = \frac{\hat{Y}_t - \hat{Y}_c}{\sqrt{\frac{\left(\sum_{j=1}^{2}\left(\sum y_i - \frac{(\Sigma x_i y_i)^2}{\Sigma x_i^2}\right)\right)\left(\frac{N}{N_t \cdot N_c} + \sum_{j=1}^{2}\frac{(\bar{X}'-\bar{X}_j)^2}{\Sigma x_i^2}\right)}{(N-4)}}} \qquad (4)$$

or

$$t_{N-4}^2 = \frac{(\hat{Y}_t - \hat{Y}_c)^2}{\frac{\left(\sum_{j=1}^{2}\left(\sum y^2 - \frac{(\Sigma x_i y_i)^2}{\Sigma x^2}\right)\right)\left(\frac{N}{N_t \cdot N_c} + \sum_{j=1}^{2}\frac{(\bar{X}'-\bar{X}_j)^2}{\Sigma x^2}\right)}{(N-4)}} \qquad (5)$$

which is equivalent to

$$t_{N-4}^2 = \frac{(\hat{Y}_t - \hat{Y}_c)^2}{PQ/R} \qquad (6)$$

where

$$P = (N_c-1)s_{cY}^2(1-r_c^2) + (N_t-1)s_{tY}^2(1-r_t^2)$$

$$Q = N \Big/ (N_t \cdot N_c) + z_t^2 \Big/ (N_t-1) + z_c^2 \Big/ (N_c-1)$$

$$R = (N-4).$$

For the regression projection model, the statistical test as given in Equation 3 is

$$t_{N-3} = \sqrt{\frac{(N_t/N)^2(\bar{Y}_t-\hat{\bar{Y}}_t)^2(N-3)}{b_c^2\bar{s}_X^2 + \bar{s}_Y^2 - 2b_c\bar{b}s_X^2 + N_tN_c/N\ (\bar{Y}_t-\hat{\bar{Y}}_t)}} \qquad (7)$$

Tallmadge and Horst (1974) demonstrated that Equation 7 is algebraically equivalent to

$$t_{N-3} = \sqrt{\frac{(N_t/N)^2(\bar{Y}_t-\hat{\bar{Y}}_t)^2(N-3)}{s_Y^2 + r_c^2 s_Y^2 - 2r_c r s_Y^2}} \qquad (8)$$

or

$$t_{N-3}^2 = \frac{(\bar{Y}_t - \hat{\bar{Y}}_t)^2}{s_Y^2(1 + r_c^2 - 2r_c r)(N/N_t)^2 / (N-3)} \qquad (9)$$

If the effects of the treatment are constant for all levels of X, the slopes of the regression lines for the treatment and comparison groups will be the same, that is, $\beta_t = \beta_c$. Under this condition,

$$E(Y_t|X') - E(Y_c|X') = (\alpha_t + \beta_t X') - (\alpha_c + \beta_c X') = \alpha_t - \alpha_t. \qquad (10)$$

Also,

$$E(\bar{Y}_t) = \sum_{i=1}^{N_t} \frac{(\alpha_t + \beta_t X_i)}{N_t} = \alpha_t + \beta_t \bar{X}_t \qquad (11)$$

and

$$E(\hat{\bar{Y}}_t) = \alpha_c + \beta_c \bar{X}_t \qquad (12)$$

because $E(\bar{Y}_t)$ is the predicted value of $\bar{X}_t$ without treatment. Therefore,

$$E(\bar{Y}_t) - E(\hat{\bar{Y}}_t) = (\alpha_t+\beta_t\bar{X}_t) - (\alpha_c+\beta_c\bar{X}_t) = \alpha_t - \alpha_c. \qquad (13)$$

From Equations 10 and 13,

$$E(Y_t|X') - E(Y_c|X') = E(\bar{Y}_t) - E(\hat{\bar{Y}}_t) \tag{14}$$

If one keeps in mind the difference in degrees of freedom of the two tests, a comparison of Equations 6 and 9 gives an approximate answer to the question of which test is more precise. The test yielding the larger value of t^2, adjusted for the different degrees of freedom, is the more precise.

By Equation 14, the numerators of Equations 6 and 9 are equal. The larger value of t^2 will result from the equation with the smallest denominator. Therefore, a comparison of the statistical tests of these two evaluation models reduces to a comparison of these denominators with an adjustment for degrees of freedom.

From Equation 6, the regression discontinuity denominator is

$$\frac{\left[(N_c-1)s_{cY}^2(1-r_c^2)+(N_t-1)s_{tY}^2(1-r_t^2)\right]\left[\frac{N}{N_tN_c} + \frac{z_t^2}{N_t-1} + \frac{z_c^2}{N_c-1}\right]}{(N-4)} \tag{15}$$

and from Equation 9 the regression projection denominator is

$$s_Y^2(1 + r_c^2 - 2r_cr)\left(\frac{N^2}{N_t^2}\right) \Big/ (N-3). \tag{16}$$

The values of r_c, r_t, s_{cY}^2, and s_{tY}^2 will vary from sample to sample. However, estimates of these values can be obtained using formulae provided by Horst (1966, pp. 414-415).

$$\rho_c^2 = \rho^2\left(\frac{\sigma_{cX}^2}{\sigma_X^2}\right) \Big/ \left(1 - \rho^2\left(1 - \frac{\sigma_{cX}^2}{\sigma_X^2}\right)\right)$$

$$\rho_t^2 = \rho^2 \left(\frac{\sigma_{tX}^2}{\sigma_X^2} \right) \Bigg/ \left(1 - \rho^2 \left(1 - \frac{\sigma_{tX}^2}{\sigma_X^2} \right) \right)$$

$$\sigma_{cY}^2 = \sigma_Y^2 \left(1 - \rho^2 \left(1 - \frac{\sigma_{cX}^2}{\sigma_X^2} \right) \right)$$

$$\sigma_{tY}^2 = \sigma_Y^2 \left(1 - \rho^2 \left(1 - \frac{\sigma_{tX}^2}{\sigma_X^2} \right) \right)$$

Substituting these quantities and simplifying, Expression 15 becomes

$$\left[(N-2) s_Y^2 (1-r^2) \right] \left[\frac{N}{N_t N_c} + \frac{z_t^2}{N_t - 1} + \frac{z_c^2}{N_c - 1} \right] \Bigg/ \left[N-4 \right]. \qquad (17)$$

Because these denominators contain different terms, the magnitude of the difference between them depends on the following: the variances of X and Y for total group and each subgroup; the correlation between X and Y for the total group and each subgroup; the z value of the cutoff score for each subgroup; and the sample sizes of each subgroup.

The value r_c is an estimate of ρ_c. Therefore,

$$\rho_c = \sqrt{\rho^2 \left(\frac{\sigma_{cX}^2}{\sigma_X^2} \right) \left(1 - \rho^2 \left(1 - \frac{\sigma_{cX}^2}{\sigma_X^2} \right) \right)}$$

was used for r_c.

A FORTRAN computer program was used to generate comparisons of Expressions 16 and 17 under the following conditions:

r^2 = .3, .4, .5, .6, .7, .8, .9

s_{cX}^2 = proportion of s_X^2 = .1, .2, .3, .4, .5

$z_{X'c}$ = 1.5, 2.0, 2.5

$z_{X't}$ = 1.5, 2.0, 2.5

N_c = 10, 15, 25, 50, 100, 200, 500

N_t = 10, 15, 25, 50, 100, 200

$.05N_c \leq N_t \leq N_c$

s^2_Y was held constant at 100 in both denominators

The program and complete results of the computer-generated comparisons of the regression projection and regression discontinuity models are available from the authors. Neither test was found to be more powerful in all situations. There were, however, obvious trends in the results. These trends are presented in Tables 1, 2, 3, and 4. The ratio of the sample sizes of the two groups proved to be more important than the actual sample size of each group. The tables are presented in terms of these ratios.

From Table 1, it can be seen that the regression projection model always provided the more precise test when the sample sizes of the two groups were equal or $N_t/N_c = 1$. In this situation, the regression projection test is more powerful because the t statistic for the regression discontinuity is smaller in absolute value than the t statistic for the regression projection test and because the rejection region for the regression projection test contains that for the regression discontinuity test, because the degrees of freedom for the regression projection test is one more than those for the regression discontinuity test.

Table 1

A Comparison of the Precision of the Regression Projection and Regression Discontinuity Models Over All Sample Sizes

		N_t/N_c					
$z_{X'c}$	$z_{X't}$	1.00	.50	.40	.25	.20	.10
1.5	1.5	RP	*	*	RD	RD	RD
2.0	1.5	RP	RP	*	RD	RD	RD
2.5	1.5	RP	RP	RP	*	RD	RD
1.5	2.0	RP	RP	RP	*	RD	RD
2.0	2.0	RP	RP	RP	*	RD	RD
2.5	2.0	RP	RP	RP	*	*	RD
1.5	2.5	RP	RP	RP	*	*	RD
2.0	2.5	RP	RP	RP	RP	*	RD
2.5	2.5	RP	RP	RP	RP	*	RD

RP = Regression Projection Model provides the more precise test

RD = Regression Discontinuity Model provides the more precise test

* = More precise test depends on the correlation and truncation in X variance; see Table 2, 3, or 4.

$z_{X'c}$ = z of the cutoff score for the comparison group

$z_{X't}$ = z of the cutoff score for the treatment group

The regression discontinuity model always provided the more precise test when the treatment group sample size was less than or equal to 10% of the comparison group sample size. In between these extremes the results were mixed and, therefore, the interpretation depends on the squared correlation coefficients and the amount of truncation in X variance. Tables 2, 3, and 4 provide this breakdown. The reader may choose the appropriate column and row for his/her sample size proportion and squared correlation coefficient, find the z scores corresponding to the cutoff score, and then determine the more precise test in each situation.

The use of these tables permits an evaluator to select the model which provides the more precise statistical test for a particular situation.

Both the regression discontinuity and regression projection evaluation models assume that the regression models of the treatment and comparison groups have equal slopes. A comparison of these two evaluation models under the condition of equal slopes has just been presented. Sometimes the effects of educational programs are not constant for all subjects. The treatment may interact with the selection variable. That is, the treatment may have an effect for some subjects but not for all, as in Figures 3 and 4.

Johnson and Neyman (1936) presented a technique which may be useful for determining program impact when the effects are not constant across all values of the selection variable. This procedure was designed to determine for which values of a control variable the difference between two groups is significant. It

Table 2

A Comparison of the Precision of the Regression Projection and Regression Discontinuity Models for $N_t/N_c = .50$ and $N_t/N_c = .40$

$(s^2_{Xt})/(s^2_X) > .1$		RP				
$(s^2_{Xt})/(s^2_X) \leq .1$						
		N_t/N_c = .50		N_t/N_c = .40		
		r^2		r^2		
$z_{X'c}$	$z_{X't}$	$\leq .7$	.8, .9	$\leq .4$	.5,.6	.7,.8,.9
1.5	1.5	RP	RD	RP	RD	RD
2.0	1.5	RP	RP	RP	RP	RD
2.5	1.5	RP	RP	RP	RP	RP
1.5	2.0	RP	RP	RP	RP	RP
2.0	2.0	RP	RP	RP	RP	RP
2.5	2.0	RP	RP	RP	RP	RP
1.5	2.5	RP	RP	RP	RP	RP
2.0	2.5	RP	RP	RP	RP	RP
2.5	2.5	RP	RP	RP	RP	RP

Table 3

A Comparison of the Precision of the Regression Projection and Regression Discontinuity Models for $N_t/N_c = .25$

		r^2						
$z_{X'c}$	$z_{X't}$	.3	.4	.5	.6	.7	.8	.9
1.5	1.5	RD	RD	RD	RD	RD	RD	RD
2.0	1.5	RD	RD	RD	RD	RD	RD	RD
2.5	1.5	*≤.4 RD *≥.5 RP	RD	RD	RD	RD	RD	RD
1.5	2.0	RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP	*≤.2 RD *≥.3 RP	*≤.2 RD *≥.3 RP	*≤.2 RD *≥.3 RP	*≤.2 RD *≥.3 RP
2.0	2.0	RP	RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP
2.5	2.0	RP	RP	RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP	*≤.1 RD *≥.2 RP
1.5	2.5	RP	RP	RP	RP	RP	RP	RP
2.0	2.5	RP	RP	RP	RP	RP	RP	RP
2.5	2.5	RP	RP	RP	RP	RP	RP	RP

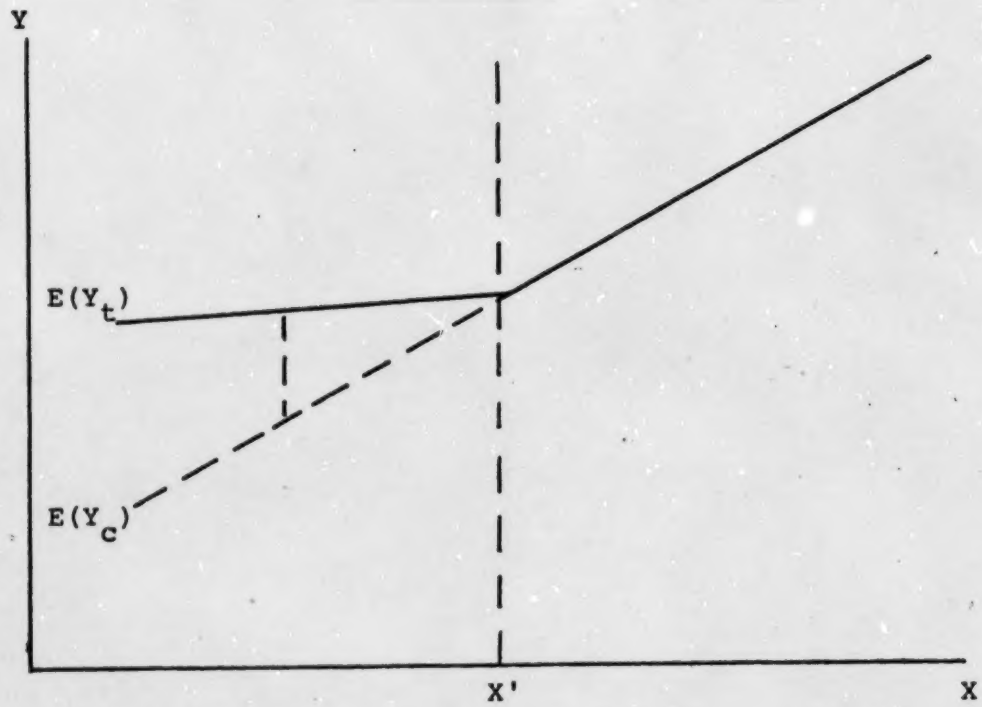
* = s^2_{Xt} / s^2_X

Table 4

A Comparison of the Precision of the Regression Projection and Regression Discontinuity Models for $N_t/N_c = .20$

$z_{X'c}$	$z_{X't}$	r^2 .3	.4	.5	.6	.7	.8	.9
1.5	1.5	RD	RD	RD	RD	RD	RD	RD
2.0	1.5	RD	RD	RD	RD	RD	RD	RD
2.5	1.5	RD	RD	RD	RD	RD	RD	RD
1.5	2.0	RD	RD	RD	RD	RD	RD	RD
2.0	2.0	RD	RD	RD	RD	RD	RD	RD
2.5	2.0	*<.2 RD *≥.3 RP	*<.3 RD *≥.4 RP	*<.3 RD *≥.4 RP	*<.3 RD *≥.4 RP	*<.3 RD *≥.4 RP	*<.3 RD *≥.4 RP	*<.3 RD *≥.4 RP
1.5	2.5	RP	RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RP
2.0	2.5	RP	RP	RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RP
2.5	2.5	RP	RP	RP	RP	*<.1 RD *≥.2 RP	*<.1 RD *≥.2 RD	*<.1 RD *≥.2 RD

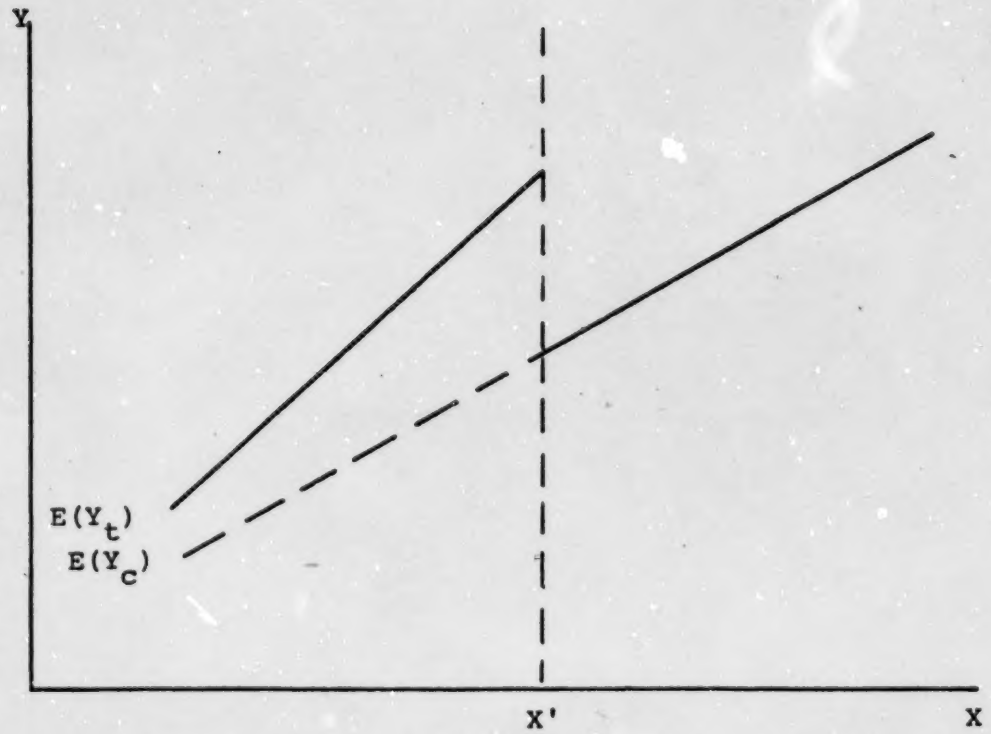
* = s^2_{Xt}/s^2_X



X = selection variable Y = dependent variable

X' = cutoff score

Figure 3. Treatment impact for participants whose selection variable score is below a certain point.



X = selection variable Y = dependent variable

X' = cutoff score

Figure 4. Treatment impact for participants whose selection variable score is above a certain point.

has become known as the Johnson-Neyman technique and has been described in several sources (Johnson & Fay, 1950; Johnson & Jackson, 1959; Walker & Lev, 1953). The technique is based on a test of $\hat{Y}_t - \hat{Y}_c$ at X':

$$\frac{\hat{Y}_t - \hat{Y}_c}{s_{\hat{Y}_t - \hat{Y}_c}} \tag{18}$$

is distributed as a t for each value of X. Therefore, the values of X for which

$$\left[\frac{\hat{Y}_t - \hat{Y}_c}{s_{\hat{Y}_t - \hat{Y}_c}}\right]^2 > t^2_{\alpha/2} \tag{19}$$

provide a two-sided region of significance such that $\hat{Y}_1 - \hat{Y}_2$ will be significant for any X within this range (Johnson & Fay, 1950, p. 351; Johnson & Neyman, 1936, p. 78).

Potthoff (1964) suggested that the Johnson-Neyman technique may be too liberal in regard to a Type I error. That is, a statement about a significant difference for all X's within a region involves a multiplicity problem that increases the probability of a Type I error. He recommended a modification that produces a more conservative region, a subset of the region obtained with the Johnson and Neyman technique.

Remembering that

$$E(Y_t) = \alpha_t + \beta_t X_i, \text{ and} \tag{20}$$

$$E(Y_c) = \alpha_c + \beta_c X_i, \tag{21}$$

a Scheffé interval for $E(Y_t|X') - E(Y_c|X')$ at any X' may be obtained. Then all the X* such that $\hat{Y}_t^* - \hat{Y}_c^*$ is outside of the Scheffe interval at X* provide a region of significance; that is, a region of X's such that a treatment effect may be inferred at all X's in that region. This region indicates a treatment effect which may be hypothetical, depending on the relationship between X' and the region.

Using methods described in Seber (1977), the Scheffé interval may be constructed in the following way (cf. Potthoff, 1964). Rewrite Equations 20 and 21 as

$$Y = X\beta + \varepsilon$$

where

$$Y = \begin{pmatrix} Y_{t1} \\ \vdots \\ Y_{tN_t} \\ Y_{c1} \\ \vdots \\ Y_{cN_c} \end{pmatrix} \quad X = \begin{pmatrix} 1 & X_{t1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{tN_t} & 0 & 0 \\ 0 & 0 & 1 & X_{c1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & X_{cN_c} \end{pmatrix} \quad \beta = \begin{pmatrix} \alpha_t \\ \beta_t \\ \alpha_c \\ \beta_c \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_{t1} \\ \vdots \\ \varepsilon_{tN_t} \\ \varepsilon_{c1} \\ \vdots \\ \varepsilon_{cN_c} \end{pmatrix}$$

and ε is multinormal with zero means and dispersion $\sigma^2 I$, I the NXN identity matrix. The estimate β is obtained through the usual least squares methods. Seber (1977, p. 129) gives the Scheffé bounds on all $h'\hat{\phi} = h'A\hat{\beta}$ as

$$h'\phi \pm (dF^{\alpha}_{d,N-p})^{\frac{1}{2}}\hat{\sigma}h'\hat{\phi} \quad , \tag{23}$$

where

$$h' = (1, X)$$

$$A = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix} \qquad \phi = A\beta,\ \hat{\phi} = A\hat{\beta},$$

d = 2 for two groups, p = 4 for four parameters, and

$$\begin{aligned} \hat{\text{var}}(h'\hat{\phi}) &= h'\hat{\mathcal{D}}[\hat{\phi}]h \\ &= h'A\hat{\mathcal{D}}[\hat{\beta}]A'h \\ &= (1\ X\ -1\ -X)\hat{\mathcal{D}}[\hat{\beta}]\begin{pmatrix} 1 \\ X \\ -1 \\ -X \end{pmatrix} \end{aligned} \tag{24}$$

The hypothesis, H_o: $A\beta = \phi = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, is accepted if and only if $(1, X)\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 0$ is in $h'\hat{\phi} \pm (2F^{\alpha}_{2,N-4})^{\frac{1}{2}}[\hat{\text{var}}(h'\hat{\phi})]^{\frac{1}{2}}$ for all h. Therefore, H_o is rejected if and only if the interval excludes 0 for some $h' = (1, X)$.

Find X such that Interval 23 excludes 0. The estimated variance of $h'\hat{\phi}$ is

$$\hat{\text{var}}(h'\hat{\phi}) = \hat{\mathcal{D}}[h'\hat{\phi}] = h'A\hat{\mathcal{D}}[\hat{\beta}]A'h,$$

where $\hat{\mathcal{D}}(\hat{\beta}) = \hat{\sigma}^2[(X'X)^{-1}]$,

$$\hat{\sigma}^2 = RSS/(N-p),$$

$$(X'X)^{-1} = \begin{pmatrix} \begin{pmatrix} d_{11} d_{12} \\ d_{21} d_{22} \end{pmatrix} & 0 \\ 0 & \begin{pmatrix} e_{11} e_{12} \\ e_{21} e_{22} \end{pmatrix} \end{pmatrix} \tag{25}$$

$$\begin{pmatrix} d_{11} d_{12} \\ d_{21} d_{22} \end{pmatrix} = \left(\frac{1}{\Sigma x_t^{\,2}}\right)\begin{pmatrix} \frac{\Sigma X_{ti}^{\,2}}{N} & -\bar{X}_t \\ -\bar{X}_t & 1 \end{pmatrix} \tag{26}$$

and

$$\begin{pmatrix} e_{11} e_{12} \\ e_{21} e_{22} \end{pmatrix} = \left(\frac{1}{\Sigma x_c^{\,2}}\right)\begin{pmatrix} \frac{\Sigma X_{ci}^{\,2}}{N} & -\bar{X}_c \\ -\bar{X}_c & 1 \end{pmatrix} \tag{27}$$

Interval 23 becomes

$$a_t + b_t X - a_c - b_c X \pm \Delta \tag{28}$$

where

$$\Delta = (2F^{\alpha}_{2,N-4})^{\frac{1}{2}} \left(\frac{RSS}{N-4}\right)^{\frac{1}{2}} \left[(d_{11}+d_{21}X,\ d_{12}+d_{22}X,\ -e_{11}-e_{21}X,\ -e_{12}-e_{22}X) \begin{matrix} 1 \\ X \\ -1 \\ -X \end{matrix}\right]^{\frac{1}{2}} \tag{29}$$

which becomes

$$(2F^{\alpha}_{2,N-4})^{\frac{1}{2}} \left(\frac{RSS}{N-4}\right)^{\frac{1}{2}} [d_{11}+d_{21}X+d_{12}X+d_{22}X^2+e_{11}+e_{21}X+e_{12}X+e_{22}X^2]^{\frac{1}{2}} \tag{30}$$

This gives the Scheffé interval for $h'\hat{\phi} = a_t + b_t X - a_c - b_c X$. For all X such that

$$(a_t + b_t X - a_c - b_c X)^2 \geq (2F^{\alpha}_{2,N-4}) \left(\frac{RSS}{N-4}\right) \cdot$$

$$[d_{11}+d_{21}X+d_{12}X+d_{22}X^2+e_{11}+e_{21}X+e_{12}X+e_{22}X^2] \tag{31}$$

there is a significant difference between $\hat{Y}_t$ and $\hat{Y}_c$. Inequality 31 may be rewritten as a quadratic in X:

$$X^2[(b_t-b_c)^2 - (2F^{\alpha}_{2,N-4})\left(\frac{RSS}{N-4}\right)(d_{22}+e_{22})] +$$

$$X[2(a_t-a_c)(b_t-b_c) - 2F^{\alpha}_{2,N-4}\left(\frac{RSS}{N-4}\right)(d_{21}+d_{12}+e_{21}+e_{12})] +$$

$$[(a_t-a_c)^2 - 2F^{\alpha}_{2,N-4}\left(\frac{RSS}{N-4}\right)(d_{11}+e_{11})] \geq 0. \tag{32}$$

The solution to this is

$$X = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{33}$$

where

$$A = (b_t-b_c)^2 - 2F^{\alpha}_{2,N-4}\left(\frac{RSS}{N-4}\right)\left(\frac{1}{\Sigma x_t^2} + \frac{1}{\Sigma x_c^2}\right) \tag{34}$$

$$B = 2\left[(a_t - a_c)(b_t - b_c) + 2F^{\alpha}_{2,N-4}\left(\frac{RSS}{N-4}\right)\left(\frac{\bar{X}_t}{\Sigma x_t^2} + \frac{\bar{X}_c}{\Sigma x_c^2}\right)\right] \quad (35)$$

$$C = (a_t - a_c)^2 - 2F^{\alpha}_{2,N-4}\left(\frac{RSS}{N-4}\right)\left(\frac{N}{N_t \cdot N_c} + \frac{\bar{X}_t^2}{\Sigma x_t^2} + \frac{\bar{X}_c^2}{\Sigma x_c^2}\right) \quad (36)$$

Geometrically, the solution to Inequality 32 is the X intercepts of the parabola whose quadratic expression is the left side of Inequality 32. These X intercepts give boundary points for a region of significance. If A, Equation 34, is positive, the parabola opens up and the regions of significance are those X's outside the boundary points. If A is negative, the parabola opens down and the region of significance includes those X's between the boundary points.

Inequality 32 yields the same result as the "simultaneous" region proposed by Potthoff (1964). These regions will always be proper subsets of the regions obtained using the Johnson-Neyman procedure. This technique provides a comparison of $\hat{Y}_t - \hat{Y}_c$ for all X's simultaneously. The Johnson-Neyman region is built on a comparison of $\hat{Y}_t - \hat{Y}_c$ at a particular X. This "simultaneous" region procedure provides a statistical test for use in evaluating the treatment effect of programs designed for selection-biased populations. A practical application demonstrating the use of the regression discontinuity, the regression projection, and the simultaneous region procedures is presented in the next section. The problem chosen for this application is an example of a hypothetical situation in which a treatment is given to a selection-biased population.

A random number generator supplied by a Texas Instruments calculator, Model SR-60, was used to generate the sample values (Texas Instruments, Inc., 1976). A total of 100 random X values were generated from a population normally distributed with mean 50 and standard deviation 10. The X values represented the scores of 100 individuals on a selection variable. All individuals with scores less than or equal to 40 were designated as the treatment group. The comparison group included all those with scores greater than 40.

A total of 100 random values from a population normally distributed with mean 0 and standard deviation 7 were generated for use as residuals. For the comparison group, the Y values were created by adding the randomly generated residuals to the X values. For the treatment group, an effect was first added to the X's and then the random residuals were added to create the Y's. This effect was purposely designed to be proportional to the X's to provide an example which included an interaction of treatment with the selection variable. Thus, for the purpose of this example, $Y_t = 2X - 20 + \varepsilon$ and $Y_c = X + \varepsilon$.

Applying the statistical test of the regression discontinuity model, Equation 6,

$$t^2_{N-4} = \frac{(54.84 - 41.25)^2 \cdot 96}{\left[17(122.59)(1-.91^2)+81(100)(1-.68^2)\right]\left[\frac{100}{1476} + \frac{1.02}{17} + \frac{2.43}{81}\right]}$$

$$= 23.84 \qquad t_{96} = 4.88.$$

This exceeds the critical values of $t_{.05,96} = 1.96$, leading to the conclusion that the difference $\hat{Y}_t - \hat{Y}_c$ at the cutoff is statistically significant.

Statistics necessary for calculating the treatment effect as measured by all three evaluation models are presented in Table 5. A graph of this example is shown in Figure 5.

Applying the statistical test of the regression progression model, Equation 9,

$$t^2_{N-3} = \frac{(18/100)^2(44.6 - 36.48)^2\ (97)}{.90^2(51.93)+(104.07)-2(.9)(51.93)+\left(\frac{18(82)}{10000}\right)\ (44.6 - 36.48)^2}$$

$$= 3.77 \qquad t_{97} = 1.94.$$

This is less than the critical value of $t_{.05,96} = 1.96$, leading to a conclusion of no significant impact at the .05 level.

For the simultaneous region procedure, Equations 33, 34, 35, and 36 yield

$$A = (1.027)^2 - 2(3.10)\left(\frac{4712.8}{96}\right)\left(\frac{1}{467.16} + \frac{1}{4641.3}\right) = .3374$$

$$B = 2\left[(1.027)(-28)+2(3.10)\left(\frac{4712.8}{96}\right)\left(\frac{35.2}{467.16} + \frac{52.3}{4641.3}\right)\right] = -4.77$$

$$C = (-28)^2 - 2(3.10)\left(\frac{4712.8}{96}\right)\left(\frac{100}{1476} + \frac{35.2^2}{467.16} + \frac{52.3^2}{4641.3}\right) = -223.27$$

$$X = \frac{4.77 \pm \sqrt{(-4.77)^2 - 4(.3374)(-223.27)}}{2(.3374)} = 33.75, -19.61$$

The value of A is positive, the parabola opens up, and these X intercepts provide two regions of significance, those X's below -19.61 and those X's above 33.75. The region below -19.61 is outside the range of interest. The region between 33.75 and 40 provides a region of X's such that a treatment effect may be inferred

Table 5

Sample Statistics for Hypothetical Problem

Statistic	Group		
	Treatment	Comparison	Total
N	18	82	100
$\bar{X}$	35.20	52.30	49.22
s^2_X	27.48	57.30	98.98
$\bar{Y}$	44.60	52.02	50.68
s^2_Y	122.59	100.15	111.06
r	.912	.68	
b	1.927	.90	
a	-23.20	4.80	
$z_{X'}$	1.01	1.56	
$\hat{Y}_{X'}$	54.84	41.25	
RSS	358.24	4354.56	

$X' = 40.5$ $\quad \hat{\bar{Y}}_t = 36.48$

$\bar{s}_X^2 = 51.93$ $\quad \bar{s}_Y^2 = 104.07$ $\quad \bar{b} = 1.08$

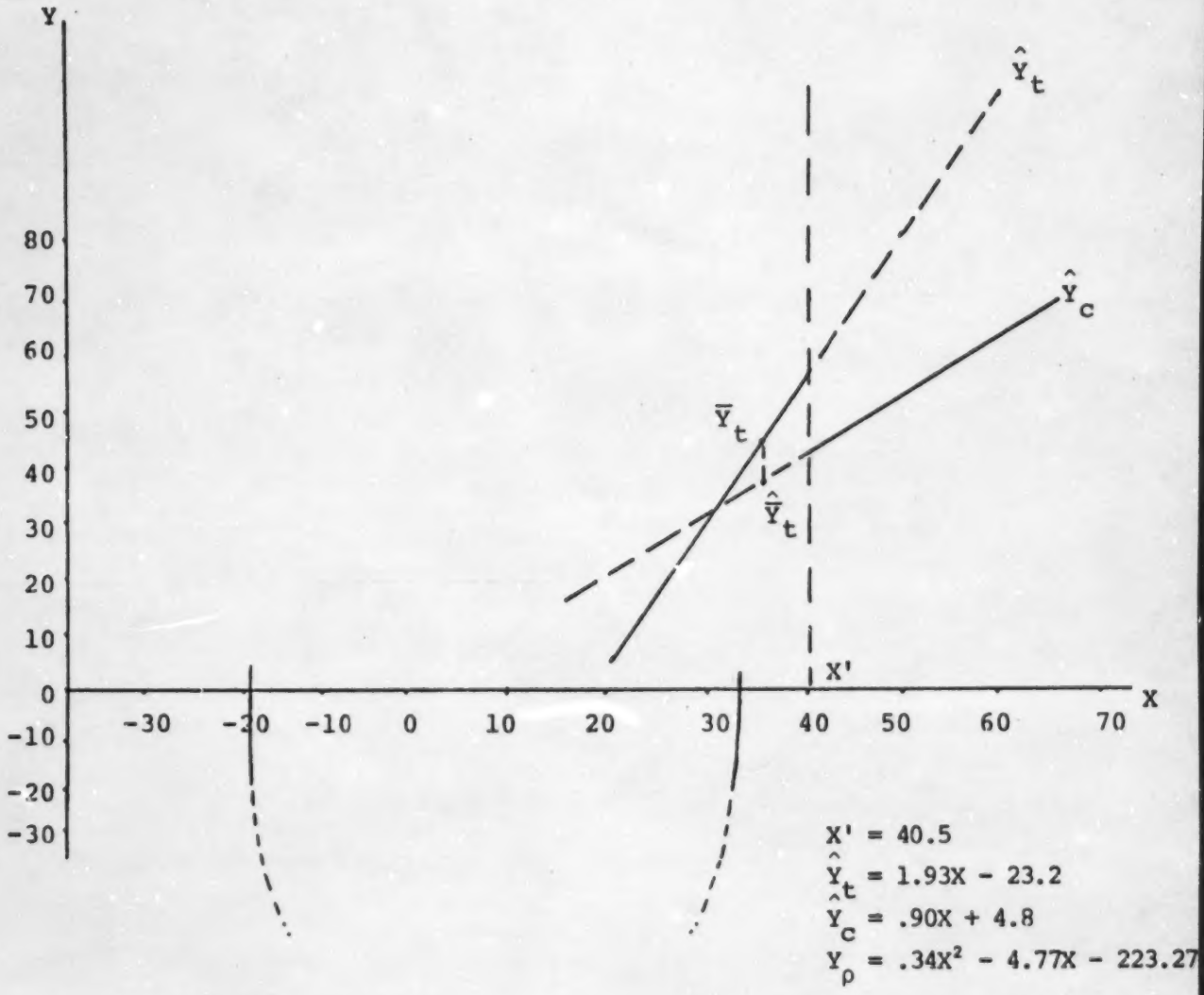


Figure 5. Graph for example problem.

for individuals whose X scores fall in this region. The region above 40 does not include any program participants. It therefore represents a region of possible treatment effect in the future use of this program if the cutoff were changed to a value above 40.

These results reaffirm the importance of using an evaluation model that determines the presence of a treatment effect by comparing the regression models at all values of a selection variable and not only at the cutoff score or treatment group mean. Using the regression discontinuity model one would infer a treatment effect when, in fact, there was none for those individuals with low scores. From the regression projection model, one would infer no treatment effect, which is not true for those program participants near the cutoff.

Using the simultaneous region technique, an evaluator may make inferences about the treatment effect at different levels of the selection variable. For example, one might conclude that there was program impact for some individuals but not for all. The program did not produce a statistically significant change for individuals with low scores on the selection variable. If the primary purpose of the program was to help these individuals, the program needs to be revised. The analysis also suggests that the program, as is, may be helpful for individuals with selection variable scores above the cutoff.

Summary and Conclusions

The purpose of this study was two-fold: (1) to compare the regression discontinuity and regression projection evaluation

models in their ability to detect a treatment effect in programs for selection-biased-populations; and (2) to determine if there was another procedure for evaluating these same programs for use when the researcher is unwilling to assume that the treatment effects are constant across all levels of the selection variable. This assumption is necessary for the correct interpretation of the results of the regression discontinuity and regression projection models.

Data collection procedures of the regression discontinuity and the regression projection models are the same. The difference in the implementation of the two models is in the test of statistical significance. Therefore, the comparison of these models could be reduced to a comparison of these significance tests to determine which test had the greater precision--that is, which test had the smaller error term or denominator in the t-test. This comparison depended on the ratio of the sample sizes of the treatment and comparison groups, the correlation between the selection variable (X) and the dependent variable (Y), the amount of truncation in the variance of the selection variable, and the z value of the cutoff score for each group. For this study 25 pairs of sample sizes, 7 XY correlations, 5 truncations in X variance, and 9 pairs of z values for a total of 7875 situations were analyzed using a FORTRAN computer program. The results of this analysis revealed that neither test was more precise in all situations. The regression projection model produced a more precise test when the sample sizes were equal or nearly equal. The regression discontinuity model test was more precise when the

sample size of the treatment group was much smaller than that of the comparison group. In between these extremes, the regression projection model was more precise when the truncation in X variance was not severe ($s_{cX}^2/s_X^2 > .2$), and the correlation between the selection variable and the dependent variable for the total population was below .70. As the truncation in the X variance became more severe and the XY correlation increased, the regression discontinuity test was more able to detect differences. Using the results listed in Tables 1, 2, 3, and 4, a user can select the appropriate or more precise test for his or her situation. However, the use of either model should be discouraged. The example problem presented showed that if an interaction between the treatment and selection variable exists, either or both of these evaluation models may lead to erroneous conclusions about program impact. To assume that an interaction does not exist may lead one to infer no treatment effect occurred or that it occurred for all participants, when in fact there may be an effect at some levels of the selection variable, but not all.

In this study, a simultaneous region procedure based on Scheffé intervals was suggested. This procedure provides a test of the difference, $\hat{Y}_t - \hat{Y}_c$, at all X's in the presence or absence of a treatment-selection variable interaction. The use of this procedure is an improvement over the regression discontinuity or regression projection models, for it allows the user to determine for which individuals--none, some, or all--a treatment effect may be inferred. This information could enable one to make recommendations about program impact, program revision, or changes in the cutoff point for admission into the program.

While neither the Johnson-Neyman nor the simultaneous region procedures are commonly used in the evaluation of enrichment or remedial programs, the Johnson-Neyman procedure has been used frequently in the evaluation of other programs in which treatment-aptitude interactions were found to exist. The Johnson-Neyman procedure is based on a difference between the two groups, $\hat{Y}_t - \hat{Y}_c$, at a particular X. Therefore, inferences about differences at all X's are made with an increased probability of a Type I error. The use of the simultaneous region procedure provides smaller regions of significance, but allows the user to make probability statements about the difference, $\hat{Y}_t - \hat{Y}_c$, for all X's simultaneously while keeping the error rate at α. For this reason the simultaneous region procedure is preferred.

It has been suggested that the Johnson-Neyman procedure need only be used after the hypothesis of parallelism or equal slopes has been rejected (Abelson, 1953). However, any test conducted following a preliminary test of parallelism is conditional on the first test and has an unknown but always higher probability of a Type I error (Olshen, 1973). For this reason, it is recommended that an evaluator proceed directly to the simultaneous region technique.

Sween (1972) recommended that preliminary tests be done to determine the underlying relationship--linear, quadratic, or cubic--between the selection and dependent variables. However, these tests also make all following tests conditional upon them with accompanying higher and unknown probability of a Type I error. Often there exists prior research on the relationship of

the variables used in these programs. Unless there is reason to believe that the relationship is nonlinear, the user should proceed directly to the simultaneous region procedure.

Concern has been expressed about the difficulty of implementing evaluation models which use regression techniques. Aside from the fact that the difficulty of a more appropriate technique should not be an issue, there is no practical basis for this concern. Hand calculators are available that perform all of the necessary statistical functions for the simultaneous region procedure. Indeed, the hypothetical example given was done entirely on a hand calculator. Large sample sizes and additional variables may make the use of high-speed computers desirable, but not necessary.

The simultaneous region, regression discontinuity, and the regression projection models are all dependent on the extrapolation of two regression lines. To the extent that there is error in this extrapolation, there may be error in the conclusions drawn from these models. However, when there are real practical pressures that preclude randomization, this model is preferable to norm comparisons, the use of matching techniques, or even no comparison at all. As in any quasi-experimental procedure, conclusions should be drawn with caution.

This study was limited to an investigation of the use of regression based evaluation models for determining program impact or treatment effect on some criterion variable. This is only one part of program evaluation. Other issues which this author did not address include goals analysis, formative or process evaluation,

needs analysis including the definition of the selection-biased population in need of the special program, the selection of the dependent and selection variables, and the reliability of these variables including the issues of test ceilings and regression effects that occur when working with extreme groups. Research is needed on these issues so that the total concept of program evaluation may be improved. Future areas of research in this area could also include studies of trends to provide evidence to support the assumption of extrapolation necessary with regression procedures; simulation studies of power to determine the ability of the simultaneous region procedure to detect treatment effects; and the application of this procedure as part of an actual program evaluation. The ultimate worth of this procedure will be determined by its ability to provide evidence of program impact with ease and confidence.

References

Abelson, R. P. A note on the Neyman-Johnson technique. Psychometrika, 1953, 18, 213-218.

Anderson, S., & Ball. S. The profession and practice of program evaluation. San Francisco: Jossey-Bass, 1978.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. Reprinted as Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Horst, D. P., Tallmadge, G. K., & Wood, C. Measuring achievement gains in educational projects. Los Altos, CA: RMC Corp., 1974 (ERIC Document Reproduction Service No. ED 114 406). Reprinted as A practical guide to measuring project impact on student achievement, 1975 (ERIC Document Reproduction Service No. ED 106 376).

Horst, P. Psychological measurement and prediction. Belmont, CA: Wadsworth, 1966.

Johnson, P. O., & Fay, L. C. The Johnson-Neyman technique, its theory and applications. Psychometrika, 1950, 15, 349-367.

Johnson, P. O., & Jackson, W. B. Modern statistical methods. Chicago: Rand McNally, 1959.

Johnson, P. O., & Neyman, J. Tests of certain linear hypotheses and their application to some educational problems. Statistical Research Memoirs, 1936, 1, 57-93.

Olshen, R. A. The conditional level of the F-test. Journal of the American Statistical Association, 1973, 68, 692-698.

Potthoff, R. F. On the Johnson-Neyman technique and some extensions thereof. Psychometrika, 1964, 29, 241-255.

Riecken, H. W., & Boruch, R. F. (Eds.) Social experimentation: A method for planning and evaluating social intervention. New York: Academic Press, 1974.

Seber, G. A. F. Linear regression analysis. New York: Wiley, 1977.

Sween, J. A. The experimental regression design: An inquiry into the feasibility of nonrandom treatment allocation (Doctoral dissertation, Northwestern University, 1972). Dissertation Abstracts International, 1972, 33, 2797B. (University Microfilms No. 72-32593)

Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Mountain View, CA: RMC Research Corporation, 1974 (ERIC Document Reproduction Service No. ED 096 344). Revised, 1976 (ERIC Document Reproduction Service No. ED 126 135)

Tallmadge, G. K., & Wood, C. T. User's guide: ESEA Title I evaluation and reporting system. Mountain View, CA: RMC Corporation, 1978.

Texas Instruments, Inc. Texas Instruments SR-60 Basic Library. Dallas, TX: Texas Instruments, Inc., 1976.

Thistlethwaite, D. L., & Campbell, D. T. Regression discontinuity analysis: An alternative to the ex post facto experiment. Journal of Educational Psychology, 1960, 51, 309-317.

Walker, H. M., & Lev, J. Statistical inference. New York: Holt, 1953.